Review for Submission 2961

This paper introduces a real-time interactive perception framework, **rt-RISeg**, designed to tackle the challenges of unseen object instance segmentation (UOIS) in robotic manipulation tasks. Departing from traditional UOIS approaches that rely heavily on large-scale training datasets and static visual cues, the proposed method instead leverages robot-induced interactions and a body frame-invariant feature (BFIF) to dynamically segment objects. Notably, the framework operates without the need for a trained segmentation model, instead utilizing relative motion cues from interactions to continuously update segmentation masks in real-time. The authors report a substantial improvement in segmentation accuracy compared to state-of-the-art UOIS methods and further highlight the potential of their approach by demonstrating its ability to generate high-quality prompts for vision foundation models.

The framework is thoughtfully constructed, with clear design decisions and comprehensive explanations that enhance the reader's understanding. The results are promising and demonstrate the potential of the method to generalize effectively across diverse manipulation scenarios. While the work is strong overall, there are a few minor issues that warrant clarification and refinement.

1 Major Comments

- There appears to be a potential issue in the computation of the effective optical flow X_t. In general, RGB-D cameras provide a projection matrix for mapping between 2D image space and 3D point clouds (referred to as $dMap$s in the paper). Assuming $dMap'_t$ represents the 3D point cloud at time t derived from I_t, the 3D velocity of points can be computed as $\Delta dMap_t = dMap'_t - dMap_t$. This velocity can then be projected back into image space using the camera's projection matrix to obtain the true E_t. This may be more accurate than directly using element-wise computation of the difference between the depth of I_{t-1} and $expectedD_t$, as currently described.

- The final portion of the IsvalidPush function is somewhat unclear. From the context, it appears the intention is to ensure that new object motions do not interfere with previously segmented objects. This might be more intuitively achieved by using $binL$ instead of $objMask$, with pixels in P masked out, to more explicitly exclude interference regions.

- Several key terms and methodologies would benefit from proper citation for the benefit of readers less familiar with the domain. Specifically: RANSAC, KMeans, Breadth-First Search, the method used to convert depth maps to point clouds, and the approach for estimating $camT_{t-1}$ from joint configurations θ_{t-1} and θ_t. If forward kinematics was used for this transformation, please clarify and cite accordingly.

- Including a brief discussion of current limitations and future directions in the conclusion would strengthen the paper. Additionally, this is a compelling project—do you plan to release it as open-source? It would also be interesting to know whether the method could be extended to handle non-rigid objects.

2 Minor Comments

- Please define the acronym SAM as Segment Anything Model.

- Since Mahalanobis distance and Markov clustering are already cited, it would be helpful to provide only high-level conceptual summaries for improved readability.

Attachment to review uploaded Mon Apr 14 10:05:22 2025